Correlación

Metodología Cuantitativa Avanzada I

Mayo 2013

Asociación y causalidad

- Técnicas de análisis causal: orientadas a detectar diferencias entre grupos
- Técnicas de asociación: orientadas a detectar similitudes entre grupos o entre características de un mismo grupo
  - En qué medidas las personas que obtienen un alto/ bajo puntaje en una medición, obtienen también un alto/bajo puntaje en otra medición?
  - En qué medida el CI de los niños se relaciona con el CI de sus padres?
    - → Pregunta original que interesaba a Spearman y Pearson (eugenesia)
- En contraste con ANOVA, ambas variables son continuas (intervalar o razón)
- En las medidas de asociación no se establece una relación de antecedente → consecuencia.

BASES: VARIANZA Y COVARIANZA

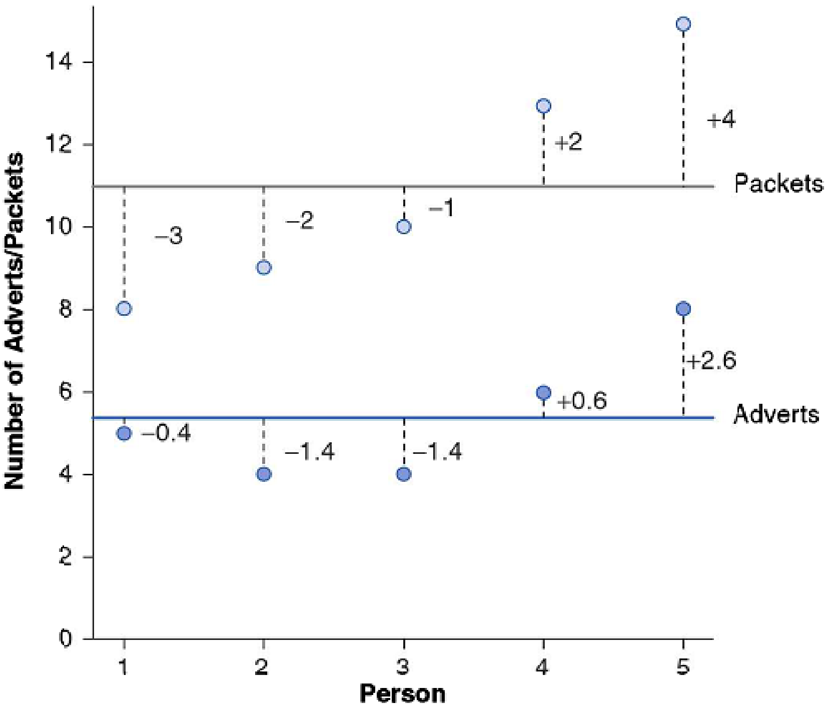
- Varianza: $$\frac{\sum (x_i - \bar{x})^2}{N-1} = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{N-1}$$

- Covarianza: $$\mathrm{cov}(\mathrm{x},\mathrm{y}) = \frac{\sum (\mathrm{x} - \bar{\mathrm{x}})(\mathrm{y} - \bar{\mathrm{y}})}{N - 1}$$

EJEMPLO VARIANZA / COVARIANZA (1)

Table 6.1 Adverts watched and toffee purchases

Participant:	*1*	*2*	*3*	*4*	*5*	*Mean*	*s*
Adverts watched	5	4	4	6	8	5.4	1.67
Packets bought	8	9	10	13	15	11.0	2.92



$$
\begin{aligned}
\mathrm{cov}(x,y) &= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{N-1} \\
&= \frac{(-0.4)(-3)+(-1.4)(-2)+(-1.4)(-1)+(0.6)(2)+(2.6)(4)}{4} \\
&= \frac{1.2+2.8+1.4+1.2+10.4}{4} \\
&= \frac{17}{4} \\
&= 4.25
\end{aligned}
$$

Fuente: Field et al (2012)

Ejemplo varianza / covarianza (2)

ID	CI_Padres	CI_hijos	x-Mx	y-My	(x-Mx)*(y-My)	(x-Mx)2	(y-My)2
1	70	89	-31.42	-14.26	448.16	987.28	203.44
2	75	73	-26.42	-30.26	799.58	698.07	915.86
3	77	79	-24.42	-24.26	592.53	596.39	588.70
4	85	86	-16.42	-17.26	283.48	269.65	298.02
5	87	87	-14.42	-16.26	234.53	207.97	264.49
6	90	91	-11.42	-12.26	140.06	130.44	150.39
7	93	98	-8.42	-5.26	44.32	70.91	27.70
8	95	102	-6.42	-1.26	8.11	41.23	1.60
9	97	102	-4.42	-1.26	5.58	19.55	1.60
10	100	110	-1.42	6.74	-9.57	2.02	45.39
11	100	111	-1.42	7.74	-10.99	2.02	59.86
12	105	116	3.58	12.74	45.58	12.81	162.23
13	110	100	8.58	-3.26	-27.99	73.60	10.65
14	113	122	11.58	18.74	216.95	134.07	351.07
15	115	105	13.58	1.74	23.58	184.39	3.02
16	120	131	18.58	27.74	515.32	345.18	769.33
17	128	115	26.58	11.74	311.95	706.44	137.75
18	130	135	28.58	31.74	907.01	816.76	1007.23
19	137	110	35.58	6.74	239.69	1265.86	45.39
Sum	1927	1962	0.00	0.00	4768	6565	5044
Prom	101.421	103.263					
N	19						

Varianza: $S^2 = \frac{\sum(X-\bar{X})^2}{n-1}$

Para CI_Padres= 6565/(19-1)

= 364.702

Para CI_hijos = 5044/19-1

280.205

Covarianza

$$\mathrm{cov}(x,y) = \frac{(x-\bar{x})(y-\bar{y})}{N-1}$$

= 4768/(19-1)

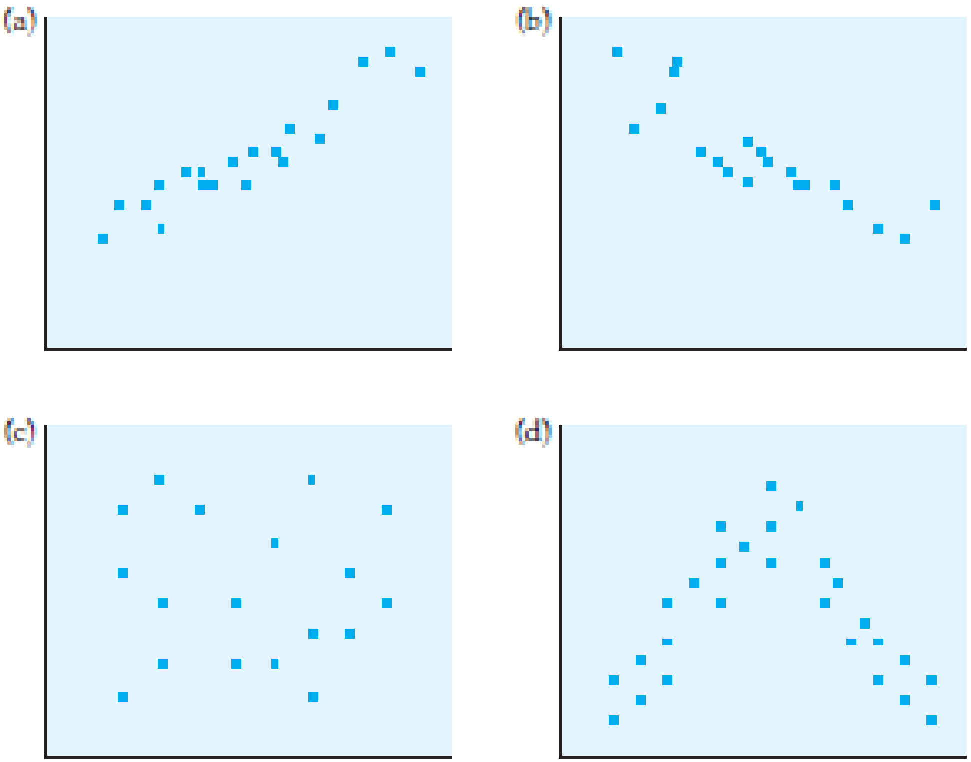
= 264.883

Correlación

- Es una medida *descriptiva*: grado de asociación lineal entre 2 variables
  - Puede ser conceptualizada como la covarianza estandarizada
  - Varía entre -1 y 1, donde 0 indica ausencia de relación (lineal) entre las variables
  - Cohen sugiere para las ciencias sociales:
    - r=0.1: correlación baja
    - r= 0.3 :correlación moderada
    - r= 0.5: correlación alta
- Es una medida *inferencial*: posee un valor probabilístico asociado con la hipótesis nula (r=0)
- La correlación no es apta para dar cuenta de relaciones no lineales, ni tampoco datos con presencia de outliers. Un diagnóstico visual se puede lograr con scatterplots (nube de puntos).

NUBE DE PUNTOS (SCATTERPLOT)

- Representación gráfica del grado de asociación entre dos variables

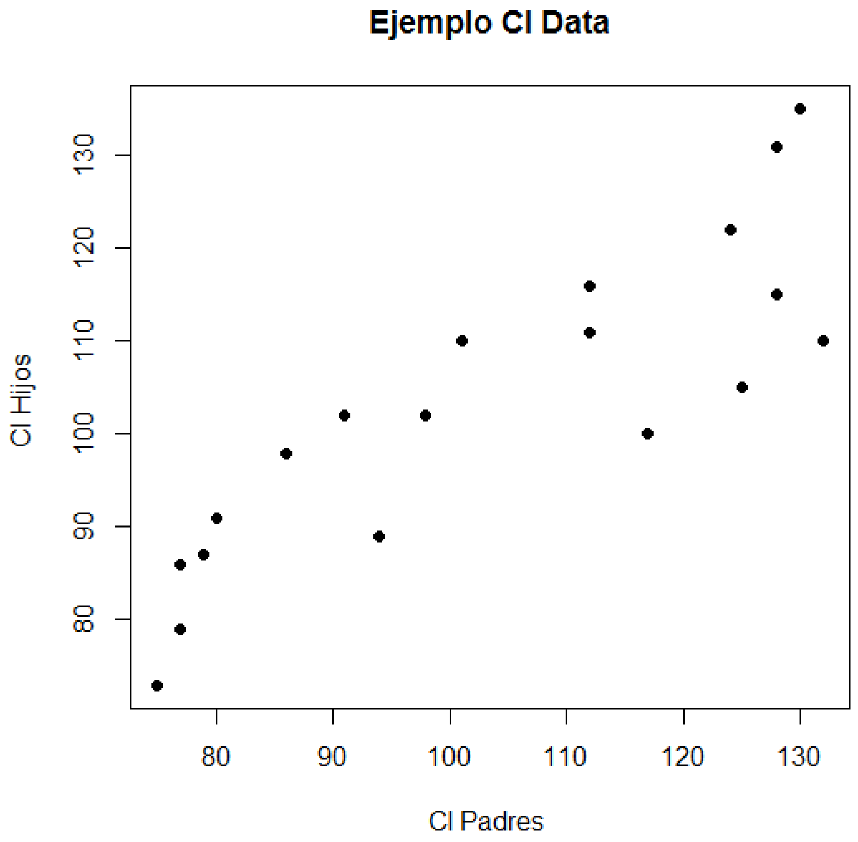


a. Positiva
b. Negativa
c. 0
d. Curvilinear

NUBE DE PUNTOS (SCATTERPLOT)

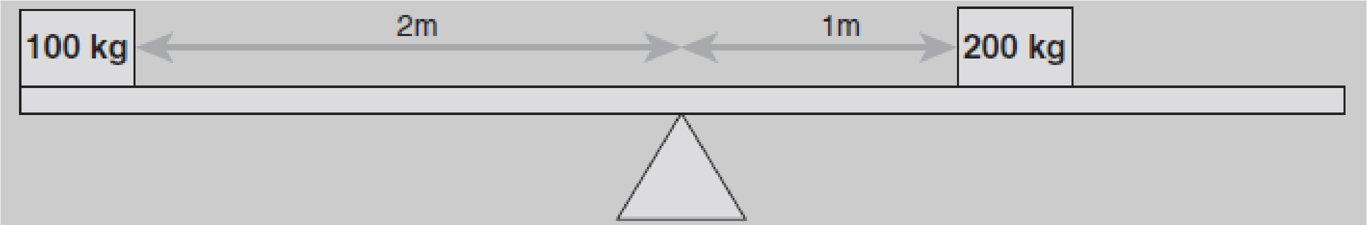
R:

plot(x,y)



Cálculo del coeficiente de correlación

- Correlación de Pearson (producto-momento)
  - Paramétrica, supone que datos son continuos y normalmente distribuidos
  - Qué es el “momento”?



  - El “momento” es igual a la distancia del centro de equilibrio multiplicado por el peso
  - En el caso de la correlación, el “centro” es el promedio. Primero se calcula la distancia del promedio para cada variable (momento), y luego ambos momentos se multiplican (producto)

CÁLCULO DEL *r* DE PEARSON

- Antecedente: Covarianza

$$\operatorname{cov}(x, y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{N - 1}$$

- Pearson:

$$r = \frac{\operatorname{cov}_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N - 1) s_x s_y}$$

- Formula computacional

$$r_{xy} = \frac{\sum[(x - \bar{x})(y - \bar{y})]}{\sqrt{\sum(x - \bar{x})^2 \times \sum(y - \bar{y})^2}}$$

Ejemplo cálculo de *r* Pearson

- CI Data

	CI_Padres	CI_hijos	x-Mx	y-My	(x-Mx)*(y-My)	(x-Mx)2	(y-My)2
	70	89	-31.42	-14.26	448.16	987.28	203.44
	75	73	-26.42	-30.26	799.58	698.07	915.86
	77	79	-24.42	-24.26	592.53	596.39	588.70
	85	86	-16.42	-17.26	283.48	269.65	298.02
	87	87	-14.42	-16.26	234.53	207.97	264.49
	90	91	-11.42	-12.26	140.06	130.44	150.39
	93	98	-8.42	-5.26	44.32	70.91	27.70
	95	102	-6.42	-1.26	8.11	41.23	1.60
	97	102	-4.42	-1.26	5.58	19.55	1.60
	100	110	-1.42	6.74	-9.57	2.02	45.39
	100	111	-1.42	7.74	-10.99	2.02	59.86
	105	116	3.58	12.74	45.58	12.81	162.23
	110	100	8.58	-3.26	-27.99	73.60	10.65
	113	122	11.58	18.74	216.95	134.07	351.07
	115	105	13.58	1.74	23.58	184.39	3.02
	120	131	18.58	27.74	515.32	345.18	769.33
	128	115	26.58	11.74	311.95	706.44	137.75
	130	135	28.58	31.74	907.01	816.76	1007.23
	137	110	35.58	6.74	239.69	1265.86	45.39
Sum	1927	1962	0.00	0.00	4768	6565	5044
Prom	101.42105	103.2632					

$$r_{xy} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\sqrt{\sum(x-\bar{x})^2 \times \sum(y-\bar{y})^2}}$$

= 4768/ RAIZ(6565 * 5044)

= **0.828605**

Alternativamente: covarXY/(sd_X*sd_Y)

covarXY=	264.883
sd_X=	19.097
sd_Y=	16.739

= 264.883/(19.097*16.739)

= 0.829

Cálculo de valor p asociado al coeficiente

- Alternativas:
  - Tabla de valores críticos p
  - Gpower
  - R

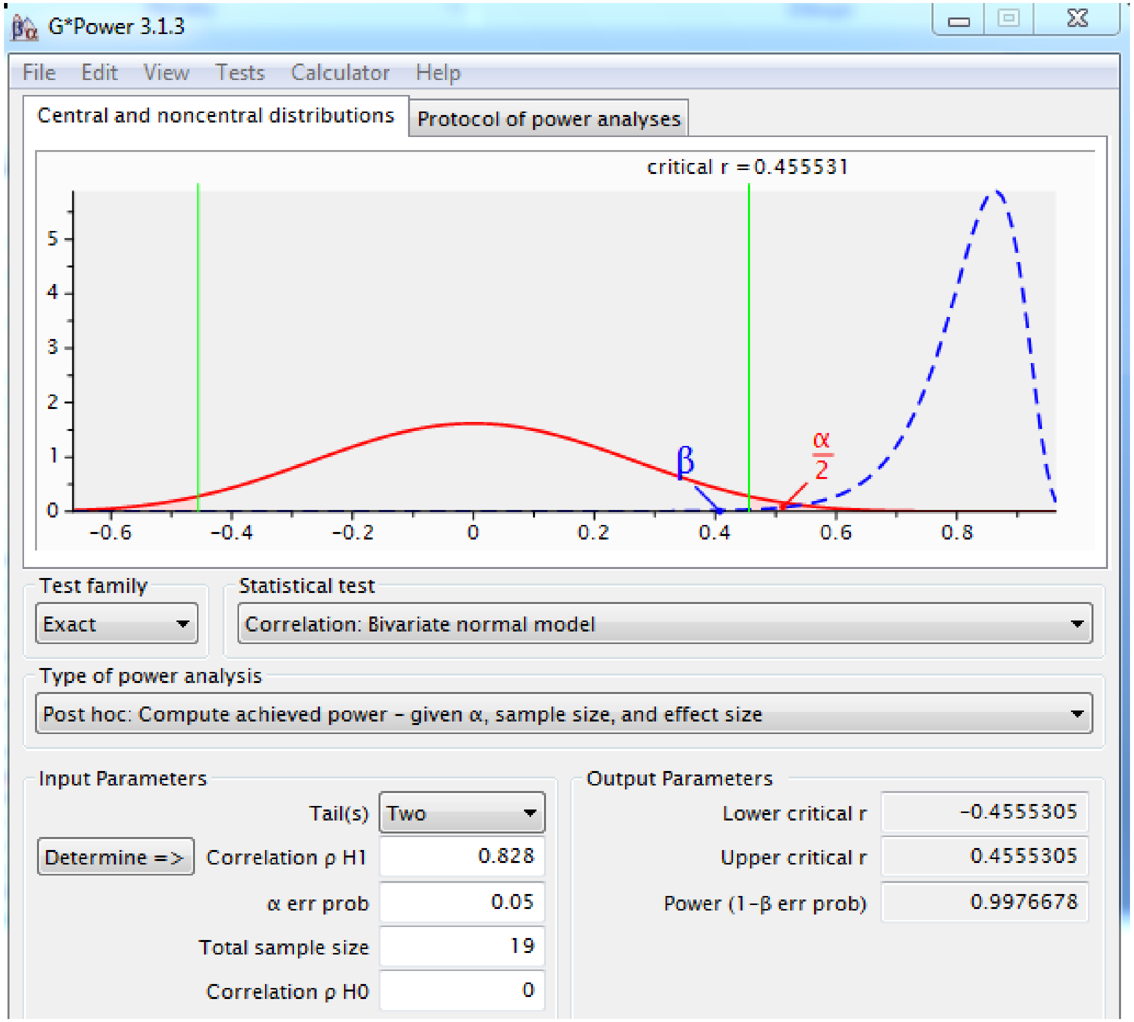
Cálculo de valor p asociado al coeficiente

- Tabla de valores críticos p

	p			
N	0.1	0.05	0.01	0.001
4	0.900	0.950	0.990	0.999
5	0.805	0.878	0.959	0.991
6	0.729	0.811	0.917	0.974
7	0.669	0.754	0.875	0.951
8	0.621	0.707	0.834	0.925
9	0.582	0.666	0.798	0.898
10	0.549	0.632	0.765	0.872
11	0.521	0.602	0.735	0.847
12	0.497	0.576	0.708	0.823
13	0.476	0.553	0.684	0.801
14	0.458	0.532	0.661	0.780
15	0.441	0.514	0.641	0.760
16	0.426	0.497	0.623	0.742
17	0.412	0.482	0.606	0.725
18	0.400	0.468	0.590	0.708
19	0.389	0.456	0.575	0.693
20	0.378	0.444	0.561	0.679
21	0.369	0.433	0.549	0.665
22	0.360	0.423	0.537	0.652
23	0.352	0.413	0.526	0.640
24	0.344	0.404	0.515	0.629
25	0.337	0.396	0.505	0.618
26	0.330	0.388	0.496	0.607

Cálculo de valor p asociado al coeficiente

- G Power



Cálculo de valor p asociado al coeficiente

- R: cor.test(x, y)

R Console

```
> cor.test(CI_padres, CI_hijos)

        Pearson's product-moment correlation

data:  CI_padres and CI_hijos
t = 6.1026, df = 17, p-value = 1.172e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6003395 0.9320339
sample estimates:
      cor
0.828605
```

Coeficientes de Correlación no paramétricos

- Para datos que no se distribuyen normalmente o que sólo son de naturaleza ordinal
- Se basan en el ranking de los datos correspondientes a los valores de cada variable
- Más usados:
  - Spearman rho
  - Kendall tau

SPEARMAN

- Basado en el ranking de los valores de la variable y en sus desviaciones
- Fórmula:

$$r = 1 - \frac{6 \times \sum d^2}{N^3 - N}$$

- El resultado de esta fórmula es cercano a una correlación de Pearson de los rankings de las variables

- Ejemplo

rating	testost	Rank rating	Rank test	d	d2
5.4	1.16	19.0	13.0	6.0	36
4.8	1.01	16.0	6.0	10.0	100
3.6	1.07	7.5	10.0	-2.5	6.25
3.8	0.81	9.5	2.0	7.5	56.25
3	1.06	3.0	8.5	-5.5	30.25
5.2	1.23	17.5	15.0	2.5	6.25
3.6	0.96	7.5	4.0	3.5	12.25
3.4	0.90	6.0	3.0	3.0	9
4.6	0.97	14.0	5.0	9.0	81
3.2	1.35	4.5	18.0	-13.5	182.25
2	0.74	1.0	1.0	0.0	0
3.2	1.14	4.5	11.0	-6.5	42.25
3.8	1.28	9.5	16.0	-6.5	42.25
4.4	1.15	12.0	12.0	0.0	0
4.6	1.34	14.0	17.0	-3.0	9
5.2	1.45	17.5	19.0	-1.5	2.25
4	1.06	11.0	8.5	2.5	6.25
2.8	1.05	2.0	7.0	-5.0	25
4.6	1.19	14.0	14.0	0.0	0
					646.5

Spearman:

$$r = 1 - \frac{6 \times \sum d^2}{N^3 - N}$$

$$= 1 - \frac{6 * 646.5}{19^3 - 19} = \mathbf{0.433}$$

Pearson rank	0.431
Pearson original	0.501

SPEARMAN VS. PEARSON

- Aplicado a los mismos datos Spearman en general tiende a otorgar valores menores (pierde información al rankear)
- La interpretación de Spearman no se realiza en términos de varianza explicada o desviaciones estándar, ya que es una técnica no paramétrica
- Lo único que se puede decir de Spearman es que muestra cuán cercanos están los rankings de dos variables

KENDALL

- Representa el grado de concordancia entre dos columnas de datos rankeados
- A mayores discrepancias, menor es el coeficiente
- Fórmula:

$$Kendall's \cdot tau = \frac{C-D}{C+D}$$

- Donde C=pares concordantes, D= pares discrepantes
- Par concordante: el número de rankings observados bajo un ranking particular que son mayores que ese ranking particular

Ejemplo Kendall

C= Cantidad de observaciones bajo esta celda que son mayores que esta celda (10)

D= Cantidad de observaciones bajo esta celda que son menores que esta celda (1)

Kendall example						
Estudiar	Ranking1	Ranking2	C	D		
a	1	2	10	1		
b	2	1	10	0		
c	3	5	7	2	Kendall =	C - D
d	4	4	7	1		C + D
e	5	3	7	0		
f	6	8	4	2	=	54 - 12
g	7	6	5	0		54 + 12
h	8	7	4	0		
i	9	12	0	3	=	**0.636**
j	10	11	0	2		
k	11	10	0	1		
l	12	9				
			54	12		

Spearman vs. Kendall

- En general, Kendall otorga coeficientes más pequeños que Spearman (por lo cual probablemente se utiliza menos)
- Se sugiere utilizar Kendall para observaciones < 20
- Kendall posee interpretación más directa que Spearman: es el grado de concordancia entre dos columnas de datos rankeados

Correlaciones con variables dicotómicas

- Correlación punto-biserial: entre una variable continua y una dicotómica

$$r = \frac{(\bar{x}_1 - \bar{x}_2) \times \sqrt{p \times (1-p)}}{\sigma_x}$$

  - Donde:
    - x_1 y x_2 son los puntajes para el grupo 1 y grupo 2,
    - p es la proporción de personas en el grupo 1, y
    - σ_x es la desviación estándar de x (ambos grupos)

Correlaciones con variables dicotómicas

- Ejemplo punto-biserial

ID	CI_Padres	CI_hijos	Sexo (1=F)			
1	70	89	1			
2	75	73	0			
3	77	79	1			
4	85	86	0			
5	87	87	1			
6	90	91	0	Prom. Mujeres=	99.600	
7	93	98	1	Prom. Hombres=	107.333	
8	95	102	0	sd(x)=	16.293	
9	97	102	1	p=	10/19 =	0.526
10	100	110	0			
11	100	111	1			
12	105	116	0	r=	99.6-107.33 * raiz(0.52 * (1-0.52)	
13	110	100	1			280.200
14	113	122	0	**r=**	**-0.237**	
15	115	105	1			
16	120	131	0			
17	128	115	1			
18	130	135	0			
19	137	110	1			
Sum	1927	1962	10			
Prom	101.421053	103.263158				

$$r = \frac{(\bar{x}_1 - \bar{x}_2) \times \sqrt{p \times (1-p)}}{\sigma_x}$$

Correlaciones con variables dicotómicas

- Ejemplo punto-biserial en R

R Console

```
> cor.test(CI_hijos, Sexo)

        Pearson's product-moment correlation

data:  CI_hijos and Sexo
t = -1.0058, df = 17, p-value = 0.3286
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6240296  0.2434180
sample estimates:
       cor
-0.2369939
```

… equivale a Pearson

Matriz de correlaciones

- Representación de asociaciones entre más de un par de variables

Matriz de correlaciones

CI_Padres	CI_hijos	PSU	Simce
94	89	319	213
75	73	341	243
77	79	361	244
77	86	363	275
79	87	365	280
80	91	372	348
86	98	439	381
91	102	446	393
98	102	453	394
101	110	480	403
112	111	524	425
112	116	585	425
117	100	589	427
124	122	614	429
125	105	625	431
128	131	629	451
128	115	630	454
130	135	681	474
132	110	790	484

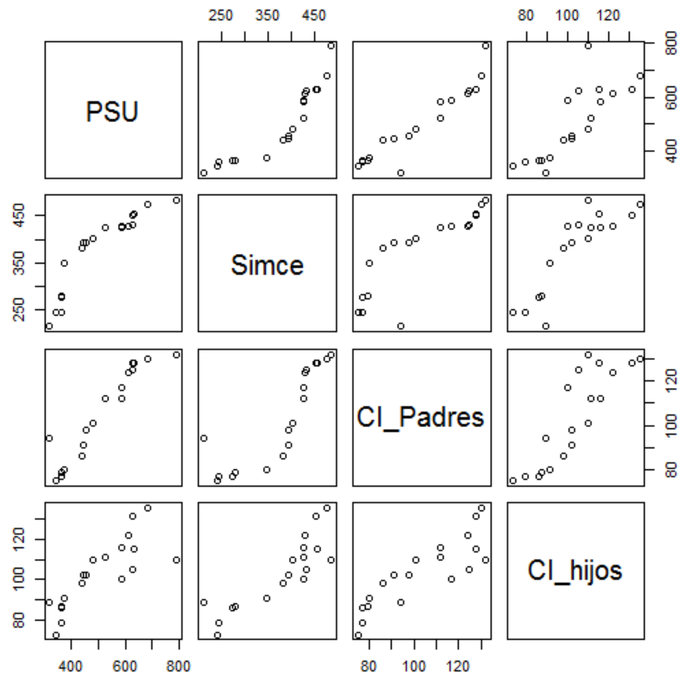
- R:

R Console

```
> cor(matrix_data)
          CI_Padres  CI_hijos       PSU     Simce
CI_Padres 1.0000000 0.8619015 0.9448937 0.8538470
CI_hijos  0.8619015 1.0000000 0.8122084 0.8638833
PSU       0.9448937 0.8122084 1.0000000 0.9053729
Simce     0.8538470 0.8638833 0.9053729 1.0000000
```

Matriz de correlaciones

- Representación gráfica en R
  - pairs(~x+y+z)
- Ej:
  - pairs(~PSU+Simce+ CI_Padres+CI_hijos)



Matriz de correlaciones

- Para obtener significación de coeficientes en matriz:
  - Por pares, ej.: cor.test(CI_padres, CI_hijos)
  - De matriz completa con comando rcor.test
    - Requiere instalar librería ltm
    - Luego: rcortest(data)

R Console

```
> rcor.test(matrix_data)

           CI_Padres CI_hijos PSU    Simce
CI_Padres   *****     0.862    0.945  0.854
CI_hijos   <0.001     *****    0.812  0.864
PSU        <0.001    <0.001    *****  0.905
Simce      <0.001    <0.001   <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

Matriz de correlaciones

- Manejo de datos perdidos
  - Listwise: se eliminan todos los casos con algún dato perdido, toda la matriz posee el mismo N
  - Pairwise: se eliminan los casos correspondientes al par de variables analizadas. El N varía por pares.

MATRIZ DE CORRELACIONES

- Ej. Datos perdidos

CI_Padres	CI_hijos	PSU	Simce
94	89	319	213
75	73	341	243
77	79	361	244
77	86	363	275
	87	365	280
80	91	372	348
86	98	439	381
91	102	446	393
98	102		394
101	110	480	403
112	111	524	425
112	116	585	425
117	100	589	427
124	122	614	429
125	105	625	
128	131	629	451
128		630	454
130	135	681	474
132	110	790	484

Matriz de correlaciones

- Listwise: cor(data, use ="na.or.complete")
- Pairwise: cor(data, use ="pairwise.complete.obs")

R Console

```
> cor(matrix_data_m, use ="na.or.complete")
          CI_Padres  CI_hijos       PSU     Simce
CI_Padres 1.0000000 0.8803434 0.9374751 0.8374678
CI_hijos  0.8803434 1.0000000 0.8118090 0.8617640
PSU       0.9374751 0.8118090 1.0000000 0.9021616
Simce     0.8374678 0.8617640 0.9021616 1.0000000
> cor(matrix_data_m, use ="pairwise.complete.obs")
          CI_Padres  CI_hijos       PSU     Simce
CI_Padres 1.0000000 0.8528051 0.9411484 0.8401963
CI_hijos  0.8528051 1.0000000 0.8081260 0.8667961
PSU       0.9411484 0.8081260 1.0000000 0.9132621
Simce     0.8401963 0.8667961 0.9132621 1.0000000
```

Matriz de correlaciones

- Significación con datos perdidos en matriz
  - rcor.test (data, + especificación de perdidos)
  - Ej:
    - rcor.test(matrix_data_m, use ="na.or.complete")
    - rcor.test(matrix_data_m, use ="pairwise.complete.obs")

Matriz de correlaciones

- Número de casos listwise / pairwise
  - Listwise: nrow(na.omit(matrix_data_m))
  - Pairwise? Alternativa:
    - Convertir datos a matriz y utilizar comando rcorr(data), que entrega matriz con N por pares
    - Ejemplo:
      - x=as.matrix(matrix_data_m)
      - rcorr(x)

CONTROL ESTADÍSTICO Y CORRELACIONES PARCIALES

Control estadístico

- ¿Es la relación que estoy estudiando lo que yo pienso que es? ¿Puede haber otra variable no considerada que afecte la relación?
- En investigación científica, control significa control de la varianza (ej. en investigación experimental)
- Control estadístico: métodos para identificar, aislar o neutralizar varianza en la variable dependiente que puede ser causada por una variable ajena al estudio
- La regresión múltiple es una de las formas que permite establecer control estadístico

Correlaciones Parciales

- *Correlación parcial*: correlación entre 2 variables (Y, X_1,) controlando el efecto de una tercera variable (X_2)

$$r_{x1y.x2} = \frac{r_{x1y} - r_{yx2} r_{x1x2}}{\sqrt{1 - r_{yx2}^2}\sqrt{1 - r_{x1x2}^2}}$$

- *Correlación semi-parcial* (part correlation): correlación entre 2 variables (Y, X_1) controlando el efecto de una tercera (X_2) solo en la variable independiente

$$r_{y(x1.x2)} = \frac{r_{x1y} - r_{yx2} r_{x1x2}}{\sqrt{1 - r_{x1x2}^2}}$$

Control mediante correlación parcial

- Se utiliza para dar cuenta de correlaciones espúreas, o también para estudiar mediación entre variables
- Ejemplo: ¿en qué medida las habilidades matemáticas y la altura correlacionan en niños, una vez controlando por (parcializando) la edad?
- El símbolo para una correlación entre 2 variables parcializando a una tercera es $r_{12.3}$
- La variable que es parcializada usualmente es llamada variable control
- Es equivalente a una correlación entre residuos entre la variables de las cuales el efecto de una tercera es parcializado

Control mediante correlación parcial

- Ejemplo

	Y	X1	X2
	1	3	3
	2	1	2
	3	2	1
	4	4	4
	5	5	5
Sum	15	15	15
Prom	3	3	3
r Yx1	0.7		
r Yx2	0.6		
r x1x2	0.9		

$$r_{x1y.x2} = \frac{r_{x1y} - r_{yx2} r_{x1x2}}{\sqrt{1 - r_{yx2}^2}\sqrt{1 - r_{x1x2}^2}}$$

$$r_{yx1.x2} = \frac{0.7 - (0.6 * 0.9)}{\sqrt{1 - 0.6^2}\sqrt{1 - 0.9^2}} = 0.46$$

- Correlación entre Y y X1, parcializando X2 de ambas variables
- Equivale a la correlación entre el residuo de la regresión de Y en X2 y el residuo de la regresión de X1 en X2

Reporte de correlaciones APA 6th

- En texto:
  - Ej: $r(55) = .49$, $p < .01$.
    - Entre paréntesis N-2 (df)

- Matriz

		(1)	(2)	(3)	(4)
1.	CI Padres	--			
2.	CI Hijos	0.85**	--		
3.	PSU	0.94**	0.81**	--	
4.	Simce	0.84**	0.87**	0.91**	--

Nota: Correlaciones de Pearson (n=15), **p.<0.01

Reporte de correlaciones APA 6th

- Ejemplo manual APA 6th

Table X

Summary of Intercorrelations, Means, and Standard Deviations for Scores on the BSS, BDI, SAFE, and MEIM as a Function of Race

Measure	1	2	3	4	*M*	*SD*
1. BSS	—	.54*	.29*	−.23*	1.31	4.32
2. BDI	.54*	—	.34*	−.14*	8.33	7.76
3. SAFE	.19*	.30*	—	−.074	47.18	13.24
4. MEIM	−.09	−.11	−.08	—	47.19	6.26
M	1.50	9.13	39.07	37.78		
SD	3.84	7.25	13.17	7.29		

Note. Intercorrelations for African American participants ($n = 296$) are presented above the diagonal, and intercorrelations for European American participants ($n = 163$) are presented below the diagonal. Means and standard deviations for African American students are presented in the vertical columns, and means and standard deviations for European Americans are presented in the horizontal rows. For all scales, higher scores are indicative of more extreme responding in the direction of the construct assessed. BSS = Beck Suicide Scale; BDI = Beck Depression Inventory; SAFE = Societal Attitudinal Familial Environmental; MEIM = Multigroup Ethnic Identity Measure. Adapted from "An Empirical Investigation of Stress and Ethnic Identity as Moderators for Depression and Suicidal Ideation in College Students," by R. L. Walker, L. R. Wingate, E. M. Obasi, and T. E. Joiner, 2008, *Cultural Diversity and Ethnic Minority Psychology, 14*, p. 78. Copyright 2008 by the American Psychological Association.
$^*p < .01$.

INTRODUCCIÓN A REGRESIÓN

- Correlación: nos habla del grado de asociación, pero no sobre qué siginifica esta asociación en la escala de las variables
- Regresión: permite establecer como aumenta(disminuye) una variable en función de otra variable
- Linea de regresión: representación de esta asociación en términos de intercepto y pendiente